

Internal consistency and compatibility of the 3Rs and 3Vs principles for project evaluation of animal research

Laboratory Animals

0(0) 1–11

© The Author(s) 2020



Article reuse guidelines:

sagepub.com/journals-permissions

DOI: 10.1177/0023677220968583

journals.sagepub.com/home/lanMatthias Eggel¹ and Hanno Würbel²

Abstract

Using animals for research raises ethical concerns that are addressed in project evaluation by weighing expected harm to animals against expected benefit to society. A harm–benefit analysis (HBA) relies on two preconditions: (a) the study protocol is scientifically suitable and (b) the use of (sentient) animals and harm imposed on them are necessary for achieving the study's aims. The 3Rs (Replace, Reduce and Refine) provide a guiding principle for evaluating whether the use of animals, their number and the harm imposed on them are necessary. A similar guiding principle for evaluating whether a study protocol is scientifically suitable has recently been proposed: the 3Vs principle referring to the three main aspects of scientific validity in animal research (construct, internal and external validity). Here, we analyse the internal consistency and compatibility of these two principles, address conflicts within and between the 3Rs and 3Vs principles and discuss their implications for project evaluation. We show that a few conflicts and trade-offs exist, but that these can be resolved either by appropriate study designs or by ethical deliberation in the HBA. In combination, the 3Vs, 3Rs and the HBA thus offer a coherent framework for a logically structured evaluation procedure to decide about the legitimacy of animal research projects.

Keywords

Project evaluation, HBA, animal research ethics, 3Rs, 3Vs

Date received: 25 May 2020; accepted: 4 October 2020

Introduction

Every year, millions of animals are used for scientific procedures to promote scientific discovery, advance human and animal health and facilitate nature conservation. Research on animals is regulated^a on the explicit understanding that it will provide important new knowledge in these domains without causing unnecessary harm to animals. Both maximising epistemic benefit and minimising harm to animals are therefore necessary conditions for legitimate animal research.

Minimising harm to research animals is promoted by the 3Rs: replace, reduce and refine.¹ The 3Rs represent a guiding principle according to which animal research is legitimate only if the study aim cannot be achieved (a) without using (sentient) animals (i.e. replace), (b) by using fewer animals (i.e. reduce) or (c) by using husbandry conditions and experimental procedures that are less harmful and more conducive

to animal welfare (i.e. refine). The 3Rs are embedded in national and international legislation and guidelines regulating the use of animals in research (e.g. the European Commission,² the US Department of Agriculture³ and the National Research Council of the National Academies⁴).

Implementation of the 3Rs is a necessary, but not sufficient, condition for legitimate animal research.

¹Institute for Biomedical Ethics and History of Medicine, University of Zurich, Switzerland

²Animal Welfare Division, Veterinary Public Health Institute University of Bern, Switzerland

Corresponding author:

Hanno Würbel, Animal Welfare Division, Veterinary Public Health Institute, University of Bern, Länggassstrasse 120, 3012 Bern, Switzerland.

Email: hanno.wuerbel@vetsuisse.unibe.ch

Unless a study produces scientifically valid and reproducible results, animals may be wasted for inconclusive research, regardless of how little harm is imposed on them. Similar to the 3Rs principle, the 3Vs principle offer a guiding principle for evaluating and promoting scientific validity in animal research.^{5,6} The 3Vs represent the three key aspects of scientific validity in animal research: construct validity, internal validity and external validity.^{5,7} Thus, the 3Vs principle serves to enhance the scientific validity of study findings in view of maximising epistemic benefit (i.e. knowledge gain). Together, the 3Vs, 3Rs and the harm–benefit analysis (HBA) form a coherent framework, which enables a logically structured procedure for evaluating the legitimacy of animal research projects (Figure 1).

This framework is based on the rule-of-law principle of proportionality, which requires that a proposed measure (here, an animal study protocol) is (a) suitable, (b) necessary and (c) reasonable for achieving its aims. The principle of proportionality is used in cases of conflict between different fundamental rights, legal interests or legal principles. In the case of animal research, these include general personal rights, integrity of life and limb, freedom of research and animal welfare.⁸

The 3Rs and 3Vs are not completely independent of each other. The 3Rs not only serve animal welfare but also help in enhancing epistemic benefit. Pain, suffering and harm are potential confounders in animal experiments (unless they are themselves targets of the

research). Minimising adverse effects on animals and promoting animal welfare therefore help improve study validity. Conversely, improving scientific validity also helps in reducing the number of animals used to establish new knowledge. Consequently, animal welfare and scientific validity are inextricably linked. Striving to maximise both scientific validity and animal welfare is generally a win–win strategy that should form the basis of every study protocol.

Much of this has been known for a long time.^{1,9–11} So, why is there so little progress in advancing the 3Rs?¹² How do we explain the high prevalence of risks of bias¹³ and poor reproducibility in animal research?¹⁴ And why do so many scientists resist adoption of refinements of experimental procedures (e.g. environmental enrichment¹⁵ and tunnel handling¹⁶) and guidelines promoting scientific rigour (e.g. ARRIVE guidelines^{17–20})?

Limited progress in advancing the 3Rs is commonly explained by a lack of dedicated funding,¹² while poor scientific validity and the ‘reproducibility crisis’ are mainly attributed to perverse incentives promoting sloppy science.²¹ However, many scientists are concerned that policies promoting scientific rigour may stifle creativity, thereby compromising scientific progress. Similarly, many scientists perceive the 3Rs as a nuisance or threat to animal science. Both sentiments are unsubstantiated. The first is in stark contrast to a large body of evidence derived from meta-research,

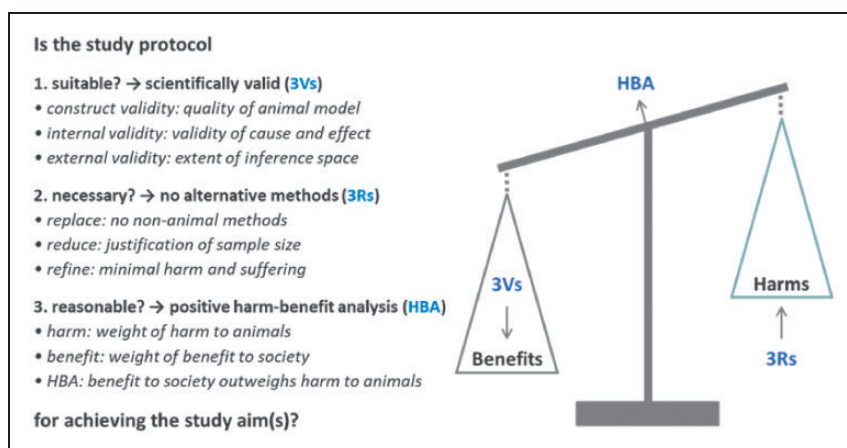


Figure 1. Framework for a structured evaluation procedure to decide about the legitimacy of animal studies. Based on the principle of proportionality, it is evaluated whether a study protocol is (a) suitable, (b) necessary and (c) reasonable for achieving the study aim(s). In the first step, suitability is determined based on the 3Vs principle. If the study protocol is deemed sufficiently suitable for achieving the study aim(s) with respect to all 3Vs (construct, internal and external validity), evaluation proceeds to the second step. In the second step, necessity is determined based on the 3Rs principle. If after application of all 3Rs the use of (sentient) animals and the harm to them are deemed necessary for achieving the study aim(s), evaluation proceeds to the third step. In the third and final step, a harm–benefit analysis (HBA) is conducted. In case of a positive HBA, that is, if the expected benefit to society (in terms of scientific discovery, human or animal health or nature conservation) is judged to outweigh the harms to animals (e.g. in terms of pain, injury or restrictions on the expression of normal behaviour), the study protocol is deemed reasonable with respect to the study aim(s). All three steps need to be met in the proposed order to justify the legitimacy of an animal study.

indicating that a lack of scientific rigour is a major source of poor reproducibility and a threat to scientific progress,²² while the second is based on a misconception of the 3Rs principle as further discussed below. Here, we examine putative conflicts between scientific and animal welfare considerations in animal research by analysing the internal consistency and compatibility of the 3Rs and 3Vs principles. We conclude that there are some conflicts and trade-offs within and between these principles, but also show how these can be resolved within the framework outlined here. Advancing both the 3Vs and the 3Rs should render animal research more useful, as well as more humane.

3Rs and 3Vs in the context of the HBA of animal research

A HBA is the common decision tool by which ethical review bodies assess the legitimacy of animal research.²³ A HBA is required by EU Directive 2010/63² and the Swiss Animal Welfare Act,²⁴ and is implied in the US Guide for the Care and Use of Laboratory Animals⁴ and the Terrestrial Animal Health Code by the World Organization for Animal Health.²⁵

In a HBA, a study's benefit to society is weighed against the harm to animals. However, as outlined above, the HBA is only the final step of a more comprehensive test of the principle of proportionality. Accordingly, two preconditions need to be met prior to a HBA, namely that for achieving the study's aim(s) (a) the study protocol is scientifically suitable, which is addressed by the 3Vs,⁵ and (b) the use of (sentient) animals and the harm inflicted on them are necessary, which is addressed by the 3Rs.^{5,26}

Whether using animals and imposing harm on them is necessary for achieving a study aim can only be reasonably assessed for a study protocol that is suitable for achieving that aim in the first place. From this follows the precedence of the 3Vs over the 3Rs in the order of the procedure of a formal HBA.⁸ This is an important procedural aspect, which clarifies that the 3Rs neither threaten animal research nor compromise its scientific validity, as assessing the 3Rs always comes secondary to assessing the 3Vs. The ultimate decision, however, as to whether a study protocol is deemed reasonable, that is, whether the harm to the animals is justified by the benefit of the study,⁵ is a moral decision (i.e. a principled judgement about the relationship between harm and benefit that leaves aside more fundamental ethical objections to invasive animal experimentation), which is taken in the HBA, after adequate implementation of the 3Vs and 3Rs has been assured.^{5,26}

Internal consistency and intra-conceptual conflicts among the 3Vs

To assess the internal consistency and identify intra-conceptual conflicts among the 3Vs, we will assess whether measures promoting one V may conflict with any of the other two Vs, and in case such conflicts exist, how they can be resolved.

Construct validity versus internal validity

Construct validity refers to the level of agreement between the animal model or outcome variable and the quality it is meant to model or measure.^{7,27–29} There is no single measure of construct validity; it is rather a judgement based on accumulated evidence, including evidence of convergent and discriminant validity.^{5,30} In contrast, internal validity refers to the extent to which the results of a given study (e.g. difference between groups or strength of relationship between variables) can be attributed to variation in the independent variable(s), rather than bias introduced by inadequacies in the design, conduct, analysis or reporting of the study.^{7,29} Similar to construct validity, there is no single measure of internal validity. It depends on appropriate controls (e.g. validated positive and negative controls) and measures to prevent bias, including randomisation, blinding, sample size calculation, as well as a priori definition of outcome variables, data handling, statistical analysis and outcome reporting.^{29,31} Thus, construct validity refers to *what* is being measured, while internal validity refers to *how* it is being measured. Conflicts or trade-offs between construct validity and internal validity are therefore unlikely to exist, and we are not aware of any such conflicts having been reported. Measures that maximise construct validity should not interfere with internal validity, and vice versa.

Construct validity versus external validity

External validity refers to the extent to which results of a given study can be applied to other studies, study conditions or animal populations (including humans^{7,29,32}). It thus defines the *inference space* of a study, that is, the range of conditions and populations to which the findings can reliably be generalised.³³ To assess whether results have high or low external validity, variation of study conditions or population characteristics is necessary.^{33,34} The current practice of rigorous standardisation in animal research entails a high risk of detecting effects with low external validity. External validity of rigorously standardised animal experiments may be so low that results may even fail to generalise to seemingly identical replicate studies.^{33–38} Measures to enhance external validity

include: splitting experiments into multiple replicate batches,^{39,40} heterogenisation of study populations by introducing systematic variation of independent variables (e.g. strain, housing conditions, test, etc.^{33,41,42}) and multi-centre study designs^{38,43}.

In principle, there should be no conflicts between construct validity and external validity. In practice, however, there may be limitations to the study of a construct under different conditions or in different populations of animals. For example, an animal model with high construct validity for a human condition may only be available in a particular mouse strain (e.g. a specific tumour model). An alternative animal model may have less construct validity but may also be available in rats and primates, thereby offering the option of a study design with higher external validity. Should we prioritise construct validity or external validity? This question may rarely occur, and the answer may not be one or the other but both. Using the model with the highest construct validity (but limited to a particular mouse strain) may be complemented by using an alternative model in additional strains or species. However, in most cases – except for regulatory toxicology, where the use of two different species (one rodent, one non-rodent) is mandatory⁴⁴ – variation of strain or species may not be necessary. It may suffice to assess external validity in terms of the findings' robustness against variation in other biologically relevant factors (e.g. age, environmental conditions, etc.^{33,40,41}).

Internal versus external validity

Experiments conducted under controlled laboratory conditions are the gold standard of study design in animal research. Defined treatment options, adequate study design, measures to prevent bias (randomisation, blinding, etc.) and stable laboratory conditions can effectively protect results against confounding. However, as indicated above, eliminating all potential confounders through rigorous standardisation carries the risk of limiting external validity of results and thus their reproducibility and generalisability.³³

Although there is no inherent conflict between internal and external validity, there may be trade-offs between measures to improve one or the other; standardisation of a study population to genetically identical animals kept under identical husbandry conditions may reduce variation in the data and thus increase precision. However, such standardisation may limit the external validity of the results to these specific standardised conditions. Because different laboratories inevitably standardise animal characteristics and environmental variables to different local study contexts, different laboratories may produce increasingly

distinct study populations as standardisation gets more rigorous.³⁶ This fruitless attempt to increase precision and reproducibility at the expense of external validity has been referred to as the standardisation fallacy.³⁴

Eliminating biological variation through standardisation is a highly inefficient strategy for generating robust evidence. Investigating each genotype–environment interaction in a separate experiment reduces information gain per experiment to virtually zero. The other extreme, however, is not an efficient strategy either. Incorporating the full range of genetic and environmental variation into every study design would render experiments unmanageable. The challenge is therefore finding the right balance between biological complexity and experimental practicability.³³

Factorial designs offer plenty of options for including biologically relevant factors (e.g. strain, sex, age or environmental parameters) into study designs as either fixed effects (for identifying sources of biological variation that modulate a treatment effect) or random effects (for estimating an average effect across a range of conditions^{33,45}). The trade-off lies in balancing the number of factors or factor levels against the number of independent replicates within each factor or factor level. Individual solutions to this trade-off depend on the intended inference space and the sources and magnitude of variation in the outcome variable. Using randomised complete block designs, internal validity can be maximised by standardisation within blocks, while external validity can be maximised by heterogenisation between blocks, whereby the heterogenisation factors should cover those factors and factor levels that determine the targeted inference space.³³

Conclusions

Taken together, there are no fundamental conflicts between the three types of scientific validity covered by the 3Vs principle, and only a few specific trade-offs exist that can be resolved by specifying the intended inference space and choosing an adequate study design. This demonstrates the high degree of intra-conceptual consistency of the 3Vs principle.

Internal consistency and intra-conceptual conflicts among the 3Rs

Several publications (e.g. Olsson et al.⁴⁶ and Boo et al.⁴⁷) discuss potential inconsistencies and intra-conceptual conflicts among the 3Rs. Olsson et al.⁴⁶ argue that 'the 3Rs are rich in ambiguities, and ... promoting one R will sometimes directly or indirectly conflict with promoting another'. Similarly, Boo et al.⁴⁷ (see also Fenwick et al.⁴⁸) concluded that 'replacement, reduction and refinement ... may have a positive or

negative effect on one or both of the other Rs'. However, we disagree with the representation of the 3Rs principle as portrayed in these publications.

The 3Rs, as conceived by Russell and Burch¹ and promoted by 3Rs centres worldwide (e.g. the UK NC3Rs <https://www.nc3rs.org.uk/> or the Swiss 3RCC <https://www.swiss3rcc.org/en/>), provide a guiding principle for performing animal research more humanely. However, they are not a strategy for eliminating animal research, for reducing the total number of animals used in research or for reducing the overall suffering imposed on them. Thus, the critique by Olsson et al.⁴⁶ that 'there is no longer any real progress in reducing the total number of animals used' or that reduction should proceed in lock step with replacement, 'as every animal test replaced by a non-animal alternative represents a reduction in the number of animals used' misses the point. Similarly, discussing putative conflicts between replacement and reduction, Boo et al.⁴⁷ miss the point when stating that 'in validation studies of replacement techniques, a comparison of the proposed new technique with the conventional in vivo technique is required, therefore having a negative impact on reduction'. This understanding of the 3Rs principle as an abolitionist strategy might explain why many scientists perceive the 3Rs as a threat to animal research.

Also, the 3Rs were never meant to replace ethical evaluation inherent to a HBA. The 3Rs serve to exploit the scope for replacement, reduction and refinement given a targeted study aim. Whether a study protocol is reasonable and thus justified for achieving a given epistemic benefit, however, needs to be determined by a HBA *after* having determined that the study protocol is suitable (e.g. according to the 3Vs) and the harm imposed on the animals is necessary (according to the 3Rs) for achieving that benefit.

This conception of the 3Rs also differs from that portrayed by both Olsson et al.⁴⁶ and Boo et al.⁴⁷ in terms of responsibility. Whereas the implementation of the 3Rs is the sole responsibility of the researcher (overseen by the competent authorities), overarching goals such as phasing out animal research, reducing total animal numbers or limiting total or maximal harm imposed on animals may be political or societal goals that are beyond the individual researcher's power and responsibility and are not part of project evaluation (although this does not mean that researchers should not be held accountable for their decisions with regards to such goals).

Despite these considerations, intra-conceptual conflicts among the 3Rs do exist and need to be resolved in view of a coherent and transparent authorisation procedure.

Replacement versus reduction

Apart from the example of additional studies required for validating new replacement methods, which, as discussed above, does not represent a true conflict, neither Boo et al.⁴⁷ nor Olsson et al.⁴⁶ have identified other conflicts between replacement and reduction, and we are not aware of any other reports of such conflicts.

Replacement versus refinement

We are not aware of any true conflicts between replacement and refinement. Use of foetal bovine serum (FBS) in cell culture studies has been discussed as a potential conflict. This would be the case if the original in vivo procedure were less harmful than the procedure for harvesting FBS. One foetus yields about 500 mL FBS,⁴⁹ which lasts for many in vitro experiments. The severity of FBS collection varies with the procedure, and according to the Foetal Calf Slaughter Welfare Protocol,⁴⁹ suffering caused by FBS collection is limited to the killing of the cow, as the calf is considered to be non-sentient before it starts breathing.⁴⁷ Thus, it seems unlikely for in vivo procedures to be less harmful than in vitro alternatives. Furthermore, as more and more serum-free media formulations for primary cell cultures and cell lines are becoming available, this potential conflict may soon disappear.

Reduction versus refinement

Reduction and refinement are the only two components among the 3Vs and 3Rs where true intra-conceptual conflicts exist. This is further complicated by a change in the meaning of the term reduction, as reflected by the recent change in the definition by the NC3Rs. In the original definition by the NC3Rs, reduction referred to 'methods which minimise the number of animals used per experiment'. Now, it also includes 'methods which allow the information gathered per animal in an experiment to be maximised in order to reduce the use of additional animals' (<https://www.nc3rs.org.uk/the-3rs>). This extension is intuitively appealing, as it places individual studies in the context of entire research programmes and focusses on maximising knowledge gain per animal rather than minimising the number of animals per study.³³ It is also consistent with Russell and Burch,¹ who state: 'One general way in which great reduction may occur is by the right choice of strategies in the planning and performance of whole lines of research'. However, methods increasing the information gathered per animal or experiment also enhance epistemic benefit. This new definition of 'reduction' may thus confound the 3Rs with the 3Vs. Although we support this adaptation for reasons mentioned above, we recommend limiting

its use to cases where more than a single experiment is evaluated. In such cases, more knowledge gained per animal in one or several experiments can be demonstrated to reduce the total number of animals used across all experiments. In all other cases, measures increasing knowledge gain should be assessed in terms of the 3Vs rather than the 3Rs.

Regardless of the definition of reduction, however, there are several examples of conflicts between reduction and refinement. These include: the use of longitudinal instead of cross-sectional studies with repeated measurements in fewer animals instead of single measurements in multiple cohorts of animals; reuse of animals for multiple experiments instead of using new animals for each experiment; and within-subjects instead of between-subjects study designs. In all of these cases, it is important to ensure that the reduction in the number of animals is balanced against additional harm to animals by the repeated or multiple application of procedures⁴⁶ (UK NC3Rs). Such trade-offs are not trivial, and we currently lack clear guidance on how to weigh numbers of animals against severity of harm in an HBA. For example, how does harm imposed by bilateral surgery on both legs of a sample of mice compare to harm imposed by unilateral surgery on only one leg of twice as many mice?

An evaluation of such trade-offs may depend on legislation (or may even be unresolvable within this kind of moral framework). For example, in contrast to German⁵⁰ and Austrian⁵¹ law, the Swiss Animal Welfare Act²⁴ does not protect the life of animals. This begs the question of how the number of killings factors into this trade-off within different jurisdictions.

Conclusions

There are no fundamental conflicts between replacement and reduction or refinement. There are, however, conflicts between reduction and refinement, resulting in specific trade-offs. In contrast to similar trade-offs between internal and external validity, however, trade-offs between reduction and refinement cannot be resolved by study design. They concern normative questions that include value judgements. Thus, the relative weight attributed to using 'more animals, each suffering less' compared to using 'fewer animals, each suffering more', depends on values (life, freedom from suffering) and judgements about these values. The responsibility for such ethical deliberation ultimately rests with the competent authorities or ethical review bodies based on a HBA and applicable law. Thus, decisions may differ between countries, depending on different norms and jurisdictions.

This further illustrates that the 3Rs principle is necessary, but not sufficient, for evaluating the legitimacy

of animal research. On the one hand, the 3Rs can only be assessed and implemented with respect to a defined epistemic benefit (i.e. the expected knowledge gain); on the other hand, legitimacy requires ethical deliberation beyond the 3Rs principles.

Compatibility of the 3Vs and 3Rs principles

In order for the 3Vs and the 3Rs to serve as complementary guiding principles in project evaluation of animal experiments, they must be compatible, that is, no fundamental conflicts must exist between them. An important component of their compatibility is the step-wise conception of our framework, which proceeds in logical steps along the three criteria of the proportionality principle (suitability, necessity and reasonableness; Figure 1). Because the 3Vs and the 3Rs concern separate questions, and because evaluation of the 3Rs comes secondary to the 3Vs, there is little scope for conflict. Furthermore, emerging conflicts or trade-offs can be resolved in the HBA, when benefits to society are weighed against harm to animals. There, the different scenarios of giving either component of a trade-off priority can be assessed in terms of their effects on the resulting harm and benefits of the study.

Given the primacy of the 3Vs in the procedure of our framework, we will now assess each of the 3Vs against the 3Rs to evaluate compatibility of the 3Vs and 3Rs and to identify potential conflicts between them.

Construct validity versus 3Rs

One of the most critical questions regarding construct validity is the model organism chosen for a study. If the target population of a study is a particular species, studying animals of that species guarantees the highest construct validity. Often, however, animals are used as model organisms for other animals, including humans, or they are used to study biological processes that apply to a range of species (e.g. all vertebrates or all mammals).

To study biological processes applicable to all animals, there is generally no justification for using sentient animals as non-sentient invertebrate model organisms are available. Fruit flies (*Drosophila melanogaster*) and nematodes (*Caenorhabditis elegans*) are the preferred model organisms to study basic biological processes, including genetics and developmental biology, but also molecular and cellular aspects of human diseases such as Parkinson's, Alzheimer's and Huntington's chorea.^{52–54}

To study biological mechanisms specific to vertebrates or mammals, the use of vertebrate or mammal model organisms is often warranted. However, for two

reasons, it is unclear whether the model organism with the highest construct validity is always used. First, the principle of 'relative replacement' is sometimes extended from non-sentient to sentient animals with a putatively lesser capacity for suffering.^{46,55} Whether the capacity for suffering varies among vertebrates (or mammals) is highly controversial from both a biological and a philosophical point of view.^{46,56–58} Nevertheless, a hierarchy is commonly applied, placing primates, dogs and cats above other mammals, mammals above birds and birds above fish. Although animal welfare legislations generally protect all vertebrates (and some invertebrates), some specific regulations exist for 'higher mammals', such as primates, dogs and cats. Such hierarchies are likely biased by (Western) human preferences for close relatives (i.e. non-human primates) and popular pets (i.e. cats and dogs) rather than based on biological evidence.⁵⁹ Nevertheless, researchers may shy away from studying these animals out of moral concerns, higher bureaucratic burden or for fear of harassment by animal rights groups. Instead of using the animal that maximises construct validity, they may choose research animals that are ethically less controversial. Second, the choice of a model organism may sometimes be based on researchers' specialisation on a particular species (e.g. mice), economic considerations (primate studies are more expensive than rodent studies) or convenience (e.g. access to facilities).

Careful evaluation of the implications of the model organism for construct validity is of great importance to the legitimacy of a study protocol. The aim to contribute to a better understanding of the pathophysiology of a human disease (e.g. Alzheimer's) does not necessarily justify the use of primates or other mammals. Some aspects of such diseases can be studied *in vitro* or by using invertebrate model organisms. It is therefore important to assess model organisms based on the specific study aims rather than the context of the study. Also, 'replacing' primates by mice for ethical reasons ('relative replacement') may miss the point if it reduces the construct validity, and thus ultimately benefit, more than the harm. More research into species differences regarding their capacity for suffering,⁵⁶ as well as research into the ethical foundation of species hierarchies, is needed to settle this issue.⁵⁷

Since there is no strong relationship between construct validity and sample size, there is little scope for conflicts between construct validity and reduction. There are, however, potential conflicts between construct validity and refinement. Although pain and suffering are potential confounders of most study outcomes, the least severe procedure may not always generate the highest construct validity. However, because construct validity can only be assessed with

respect to the specific study aim, potential refinements need to be assessed with respect to both reduction in severity of the procedures (i.e. the harm side of the HBA) and consequences for the construct validity of the study findings (i.e. the benefit side of the HBA). Consequently, conflicts between construct validity and refinement need to be addressed in the HBA.

Internal validity versus 3Rs

Besides adoption of measures to prevent risks of bias (i.e. randomisation, blinding, etc.), internal validity essentially depends on appropriate control groups (e.g. validated positive and negative controls) and an adequate sample size. Both are rather technical aspects of experimental design, for which excellent guidelines and online tools are available (e.g. Experimental Design Assistant <https://eda.nc3rs.org.uk/> and G*Power <http://www.gpower.hhu.de>). There is little potential for conflicts between internal validity and the 3Rs, except for conflicts between internal validity and reduction. This is mirrored by the UK NC3Rs' adaptation of 'reduction', which besides minimising the number of animals per study also implies 'that studies with animals are appropriately designed and analysed to ensure robust and reproducible findings'. It is clear that studies lacking essential control groups or statistical power will fail to produce sound evidence. Thus, the minimal number of animals per study should always be justified by the needs of valid inferences with respect to the expected results and the intended inference space.

As pain and suffering are potential confounders of study outcomes, refinements are likely to improve rather than compromise internal validity.^{9,10,30} Thus, rather than conflicts, there are many synergies between refinement and internal validity. However, as with construct validity, should conflicts occur, they need to be addressed by evaluating the consequences of the different scenarios for the outcome of the HBA.

External validity versus 3Rs

There are also conflicts between external validity and the 3Rs. Replacement and refinement can affect external validity only indirectly via their effects on construct validity. In contrast, there is a direct conflict between external validity and reduction. Standardisation is often promoted to reduce variation in results, thus minimising sample size to detect a treatment effect of a given size.⁶⁰ As discussed above, excessive standardisation may compromise external validity (by narrowing the inference space) and thus reproducibility and generalisability of study findings. Therefore, gains in terms of smaller sample sizes may be offset by the loss in

external validity. Factorial study designs offer plenty of opportunities to optimise study design in terms of external validity and sample size. In particular, using randomised block designs, sample size may be minimised by standardisation within blocks, while external validity may be maximised by heterogenisation between blocks.^{33,45} Specific solutions to this trade-off depend on the intended inference space and the sources and magnitude of variation of factors that determine the inference space.

Conclusions

We have shown that some conflicts and trade-offs between the 3Vs and the 3Rs exist. However, they can be resolved either by adequate study designs or by ethical deliberation in the HBA. Importantly, the 3Vs and 3Rs, together with the HBA, enable a logically structured procedure to arrive at scientifically informed and morally justified decisions about the legitimacy of animal study protocols. Careful assessment of both the 3Vs and 3Rs prior to the HBA is crucial for finding the right balance between the two. Promoting the 3Rs at the expense of the 3Vs may result in wasting animals for inconclusive research. Promoting the 3Vs at the expense of the 3Rs may result in inhumane research.

Outlook

The 3Rs and the HBA are well-established guiding principles enshrined in national and international legislation and guidelines (e.g. the European Commission,² the US Department of Agriculture³ and the National Research Council of the National Academies⁴). However, there is a lack of a similar guiding principle promoting the scientific validity of animal research. A study in Switzerland recently showed that authorities licensing animal experiments lack important information about experimental design and conduct.¹⁸ As a result, animal experiments were authorised based on blind trust into their scientific validity rather than evidence presented in study protocols (the application form has now been changed). In light of accumulating evidence of poor reproducibility in animal research, voices calling for measures to improve scientific validity are getting louder (e.g. Macleod,²¹ Stark⁶¹ and Bishop⁶²). The 3Vs offer a guiding principle for evaluating the scientific validity of animal research in the context of the HBA.⁵

The 3Vs are not the first proposal to include quality of research more formally into project evaluation. For example, Bateson's famous decision cube⁶³ included a third dimension (besides harm and benefit) – 'importance of research' – which covers scientific quality as assessed by peer review in funding decisions. Later,

Porter⁶⁴ proposed ethical scores for animal experiments, including 'that the experiment be well planned and statistically sound and seeks a realistic judgment of its exigency'. However, the 3Vs provide a much more specific and formalised principle for evaluating whether a study protocol is scientifically suitable. Moreover, the 3Vs are already operationalised and can be readily implemented into project evaluation.

More recently, Strech and Dirnagl⁶⁵ proposed a similar framework that goes even further. They propose to extend the original 3Rs by another set of three Rs to cover the scientific value of research: robustness, registration and reporting. While robustness essentially covers the 3Vs as discussed here, Strech and Dirnagl consider an obligation for both the preregistration of study protocols and the reporting of results as crucial measures to guarantee scientific value. Their proposal bears a certain risk of diluting the strong brand and coherence of the 3Rs principle. Moreover, current law and regulatory documents provide a regulatory basis for implementing the 3Vs in project evaluation. This is currently not the case for preregistration and reporting that do not have the same regulatory status in animal research. Nevertheless, we strongly support their call for preregistration and reporting as additional requirements for improving the scientific value of animal research, and hope that the necessary legal and regulatory foundations for their implementation will soon be laid.

Finally, DeGrazia and Beauchamp⁶⁶ recently proposed replacing the 3Rs principle (which they consider inadequate for evaluating animal research ethics) by three principles for animal welfare (no unnecessary harm, basic needs and upper limits to harm) and three principles for social benefit (no alternative methods, expected net benefit and sufficient value to justify harm). Unfortunately, they mistook the 3Rs principle for a framework for ethical evaluation, ignoring that ethical evaluation is based on a comprehensive HBA. Except for an 'upper limit to harm', all other 'new' principles are either covered by legal minimal standards ('basic needs'), the 3Rs ('no alternative method', 'no unnecessary harm') or by the HBA ('expected net benefit', 'sufficient value to justify harm'). Moreover, the EU Directive does actually set a limit on permissible harm,² and most other jurisdictions include some deontological (i.e. animal rights based; e.g. ban on using great apes), besides utilitarian, principles of animal ethics. In fact, rather than demonstrating a need for new principles, DeGrazia and Beauchamp argue for more rigorous implementation of the current principles. Thus, all of their principles can be accommodated within current frameworks; 'basic needs' and 'upper limits to harm' are best regulated by minimal standards in animal welfare law, 'no alternative method' and 'no

unnecessary harm' are covered by the 3Rs and 'expected net benefit' and 'sufficient value to justify harm' are accommodated by the HBA.

Taken together, we believe that the rule-of-law principle of proportionality offers an ideal basis for an effective framework of animal research ethics. We are confident that it will be sufficiently robust to stand the test of time by accommodating evolving shifts in scientific and ethical standards.

Declaration of Conflicting Interests

The author(s) declared no potential conflicts of interest with respect to the research, authorship and/or publication of this article.

Funding

The author(s) disclosed receipt of the following financial support for the research, authorship and/or publication of this article: M.E. was funded by the Messerli Foundation.

ORCID iD

Hanno Würbel  <https://orcid.org/0000-0002-2934-3010>

Note

- a. Unless stated otherwise, we refer to EU legislation throughout this article.

References

1. Russell W and Burch R. *The principles of humane experimental technique*. London: Methuen, 1959.
2. European Commission. Directive 2010/63/EU of the European Parliament and of the Council of 22 September 2010 on the protection of animals used for scientific purposes. *Off J Eur Union* 2010; 28: 82–128.
3. United States Department of Agriculture. Animal Welfare Act and Animal Welfare Regulations 2019, https://www.aphis.usda.gov/animal_welfare/downloads/bluebook-ac-awa.pdf (accessed 30 April 2020).
4. National Research Council of the National Academies. *Guide for the care and use of laboratory animals*. 8th ed. Washington, DC: National Academies Press, 2011.
5. Würbel H. More than 3Rs: the importance of scientific validity for harm–benefit analysis of animal research. *Lab Animal* 2017; 46: 164–166.
6. Sena ES and Currie GL. How our approaches to assessing benefits and harms can be improved. *Anim Welf* 2019; 28: 107–115.
7. Henderson VC, Kimmelman J, Fergusson D, et al. Threats to validity in the design and conduct of preclinical efficacy studies: a systematic review of guidelines for in vivo animal experiments. *PLoS Med* 2013; 10: e1001489.
8. German Research Foundation. *Animal experimentation in research: the 3Rs principle and the validity of scientific research – guidelines of the Permanent Senate Commission on animal protection and experimentation of the DFG for the design and description of animal experimental research projects*. Bonn: DFG, 2019.
9. Poole T. Happy animals make good science. *Lab Anim* 1997; 31: 116–124.
10. Würbel H. Ideal homes? Housing effects on rodent brain and behaviour. *Trends Neurosci* 2001; 24: 207–211.
11. Garner JP. Stereotypes and other abnormal repetitive behaviors: potential impact on validity, reliability, and replicability of scientific outcomes. *ILAR J* 2005; 46: 106–117.
12. Percie Du Sert N and Robinson V. The NC3Rs gateway: accelerating scientific discoveries with new 3Rs models and technologies. *F1000Res* 2018; 7: 591.
13. Macleod MR, McLean AL, Kyriakopoulou A, et al. Risk of bias in reports of in vivo research: a focus for improvement. *PLoS Biol* 2015; 13: e1002273.
14. Macleod M and Mohan S. Reproducibility and rigor in animal-based research. *ILAR J* 2019; 60: 17–23.
15. Toth LA, Kregel K, Leon L, et al. Environmental enrichment of laboratory rodents: the answer depends on the question. *Compar Med* 2011; 61: 314–321.
16. Henderson LJ, Smulders TV and Roughan JV. Identifying obstacles preventing the uptake of tunnel handling methods for laboratory mice: an international thematic survey. *PLoS One* 2020; 15: e0231454.
17. Kilkenny C, Browne WJ, Cuthill IC, et al. Improving bioscience research reporting: the ARRIVE guidelines for reporting animal research. *PLoS Biol* 2010; 8: e1000412.
18. Vogt L, Reichlin TS, Nathues C, et al. Authorization of animal experiments is based on confidence rather than evidence of scientific rigor. *PLoS Biol* 2016; 14: e2000598.
19. Hair K, Macleod MR, Sena E, et al. A randomised controlled trial of an Intervention to Improve Compliance with the ARRIVE guidelines (IICARus). *Res Integr Peer Rev* 2019; 4: 12.
20. The NPQIP Collaborative Group. Did a change in *Nature* journals' editorial policy for life sciences research improve reporting? *BMJ Open Sci* 2019; 3: e000035.
21. Macleod M. Why animal research needs to improve. *Nature* 2011; 477: 511.
22. Munafò M, Nosek B, Bishop D, et al. A manifesto for reproducible science. *Nat Hum Behav* 2017; 1: 0021.
23. Brønstad A, Newcomer CE, Decelle T, et al. Current concepts of harm–benefit analysis of animal experiments – report from the AALAS–FELASA Working Group on harm–benefit analysis – Part 1. *Lab Anim* 2016; 50: 1–20.
24. Swiss Animal Welfare Act (SR 455), www.admin.ch/opc/de/classified-compilation/20022103/index.html (accessed 14 July 2020).
25. World Organisation for Animal Health. Terrestrial animal health code 2019, <https://www.oie.int/standard-setting/terrestrial-code/> (accessed 27 March 2020).
26. Swiss Academies of Arts and Sciences. Weighing of interests for proposed animal experiments. Guidance for applicants. *Swiss Acad Commun* 2017; 12.
27. Cronbach LJ and Meehl PE. Construct validity in psychological tests. *Psychol Bull* 1955; 52: 281–302.

28. Shadish WR, Cook TD and Campbell DT. *Experimental and quasi-experimental designs for generalized causal inference*. Boston: Houghton, Mifflin and Company, 2002, pp.xxi, 623.
29. Bailoo JD, Reichlin TS and Würbel H. Refinement of experimental design and conduct in laboratory animal research. *ILAR J* 2014; 55: 383–391.
30. Garner J, Gaskill B, Weber E, et al. Introducing therioepistemology: the study of how knowledge is gained from animal research. *Lab Anim* 2017; 46: 103–113.
31. Vollert J, Schenker E, Macleod M, et al. Systematic review of guidelines for internal validity in the design, conduct and analysis of preclinical biomedical experiments involving laboratory animals. *BMJ Open Sci* 2020; 4: e100046.
32. Lehner PN. *Handbook of ethological methods*. Cambridge: Cambridge University Press, 1996.
33. Voelkl B, Altman NS, Forsman A, et al. Reproducibility of animal research in light of biological variation. *Nat Rev Neurosci* 2020; 21: 384–393.
34. Würbel H. Behaviour and the standardization fallacy. *Nat Genet* 2000; 26: 263–263.
35. Crabbe JC, Wahlsten DL and Dudek BC. Genetics of mouse behavior: interactions with laboratory environment. *Science* 1999; 284: 1670–1672.
36. Richter SH, Garner JP and Würbel H. Environmental standardization: cure or cause of poor reproducibility in animal experiments? *Nat Methods* 2009; 6: 257–261.
37. Kafkafi N, Golani I, Jaljuli I, et al. Addressing reproducibility in single-laboratory phenotyping experiments. *Nat Methods* 2017; 14: 462.
38. Voelkl B, Vogt L, Sena ES, et al. Reproducibility of preclinical animal research improves with heterogeneity of study samples. *PLoS Biol* 2018; 16: e2003693.
39. Paylor R. Questioning standardization in science. *Nat Methods* 2009; 6: 253–254.
40. Karp NA, Speak AO, White JK, et al. Impact of temporal variation on design and analysis of mouse knockout phenotyping studies. *PLoS One* 2014; 9: e111239.
41. Richter SH, Garner JP, Auer C, et al. Systematic variation improves reproducibility of animal experiments. *Nat Methods* 2010; 7: 167–168.
42. Richter SH, Garner JP, Zipser B, et al. Effect of population heterogenization on the reproducibility of mouse behavior: a multi-laboratory study. *PLoS One* 2011; 6: e16461.
43. Wodarski R, Delaney A, Ultenius C, et al. Cross-centre replication of suppressed burrowing behaviour as an ethologically relevant pain outcome measure in the rat: a prospective multicentre study. *Pain* 2016; 157: 2350–2365.
44. European Medicines Agency. ICH guideline M3(R2) on non-clinical safety studies for the conduct of human clinical trials and marketing authorisation for pharmaceuticals. EMA/CPMP/ICH/286/1995, https://www.ema.europa.eu/en/documents/scientific-guideline/ich-guide-line-m3r2-non-clinical-safety-studies-conduct-human-clinical-trials-marketing-authorisation_en.pdf (accessed 30 March 2020).
45. Krzywinski M and Altman N. Points of significance: analysis of variance and blocking. *Nat Methods* 2014; 11: 699–700.
46. Olsson IAS, Franco NH, Weary DM, et al. The 3Rs principle – mind the ethical gap! In: *ALTEX proceedings of the 8th world congress on alternatives and animal use in the life sciences, Montreal 2011*, pp.333–336. Baltimore: Johns Hopkins University Press, 2012.
47. Boo MJ, Rennie AE, Buchanan-Smith HM, et al. The interplay between replacement, reduction and refinement: consideration where the Three R's interact. *Anim Welf* 2005; 4: 327–332.
48. Fenwick N, Griffin G and Gauthier C. The welfare of animals used in science: how the 'Three Rs' ethic guides improvements. *Can Vet J* 2009; 50: 523–530.
49. Van Der Valk J, Mellor D, Brands R, et al. The humane collection of fetal bovine serum and possibilities for serum-free cell and tissue culture. *Toxicol In Vitro* 2004; 18: 1–12.
50. Deutsches Tierschutzgesetz, <https://www.gesetze-im-internet.de/tierschg/BJNR012770972.html> (accessed 25 April 2020).
51. Österreichisches Tierschutzgesetz, <https://www.ris.bka.gv.at/GeltendeFassung.wxe?Abfrage=Bundesnormen&Gesetzesnummer=20003541> (accessed 25 April 2020).
52. Bonini NM and Fortini ME. Human neurodegenerative disease modeling using *Drosophila*. *Annu Rev Phytopathol* 2002; 40: 381–410.
53. Iijima K, Liu HP, Chiang AS, et al. Dissecting the pathological effects of human Abeta40 and Abeta42 in *Drosophila*: a potential model for Alzheimer's disease. *Proc Natl Acad Sci U S A* 2004; 101: 6623–6628.
54. Iijima K and Iijima-Ando K. *Drosophila* models of Alzheimer's amyloidosis: the challenge of dissecting the complex mechanisms of toxicity of amyloid-beta 42. *J Alzheimers Dis* 2008; 15: 523–540.
55. Redmond C. When is an alternative not an alternative? Supporting progress for absolute replacement of animals in science. In: Herrmann K and Jayne K (eds) *Animal experimentation: working towards a paradigm change*. Vol. 22. Leiden: Brill, 2019, pp.654–672.
56. Proctor H. Animal sentience: where are we and where are we heading? *Animals* 2012; 2: 628–639.
57. Bovenkerk B and Kaldewaij F. The use of animal models in behavioural neuroscience research. In: Lee G, Illes J and Ohl F (eds) *Ethical issues in behavioral neuroscience*. Berlin: Heidelberg, 2015, pp.17–46.
58. Wild M. *Fische Kognition, Bewusstsein und Schmerz – Eine philosophische Perspektive*. Bern: EKAH, 2012.
59. Van Schaik C and Burkart J. Sind höhere Tiere mehr wert als niedere? Ein Versuch zur Exegese. In: Sigg H and Folkers G (eds) *Güterabwägung bei der Bewilligung von Tierversuchen – Die Güterabwägung interdisziplinär-kritisch beleuchtet*. Zurich: Collegium Helveticum Heft 11, 2011, pp.71–76.
60. Festing MFW. Evidence should trump intuition by preferring inbred strains to outbred stocks in preclinical research. *ILAR J* 2014; 55: 399–404.

61. Stark PB. Before reproducibility must come preproducibility. *Nature* 2018; 557: 613.
62. Bishop D. Rein in the four horsemen of irreproducibility. *Nature* 2019; 568: 435.
63. Bateson P. When to experiment on animals. *New Sci* 1986; 109: 30–32.
64. Porter DG. Ethical scores for animal experiments. *Nature* 1992; 356: 101–102.
65. Strech D and Dirnagl U. 3Rs missing: animal research without scientific value is unethical. *BMJ Open Sci* 2019; 3: bmjopen-2018-000048.
66. DeGrazia D and Beauchamp TL. Beyond the 3 Rs to a more comprehensive framework of principles for animal research ethics. *ILAR J* 2019; ilz011.

Résumé

L'utilisation des animaux pour la recherche soulève des préoccupations éthiques qui sont abordées dans l'évaluation du projet en comparant les préjudices attendus pour les animaux par rapport aux avantages attendus pour la société. Une analyse bénéfice-préjudice (HBA) repose sur deux conditions préalables, à savoir que (a) le protocole d'étude soit scientifiquement approprié et (b) que l'utilisation d'animaux (sensibles) et les préjudices qui leur sont imposés soient nécessaires pour atteindre les objectifs de l'étude. Les 3R (remplacer, réduire et raffiner) fournissent un principe directeur pour évaluer si l'utilisation des animaux, leur nombre et les préjudices qui leur sont imposés sont nécessaires. Un principe directeur similaire pour évaluer si un protocole d'étude est scientifiquement approprié a récemment été proposé: le principe 3V, qui fait référence aux trois principaux aspects de validité scientifique dans la recherche animale (construction, validité interne et externe). Nous analysons ici la cohérence et la compatibilité internes de ces deux principes, abordons les conflits à l'intérieur et entre les 3R et les 3V et discutons de leurs implications pour l'évaluation du projet. Nous montrons qu'il existe quelques conflits et compromis, mais que ceux-ci peuvent être résolus soit par des études appropriées, soit par des délibérations éthiques lors de l'analyse HBA. En combinaison, les 3V, les 3R et l'analyse HBA offrent ainsi un cadre cohérent pour qu'une procédure d'évaluation logiquement structurée permette de décider de la légitimité des projets de recherche animale.

Abstract

Die Verwendung von Tieren zu Forschungszwecken wirft ethische Bedenken auf, die bei der Projektevaluierung durch Abwiegen von zu erwartendem Schaden für die Tiere und zu erwartenden Nutzen für die Gesellschaft zu berücksichtigen sind. Eine solche Schaden-Nutzen-Analyse (Harm-Benefit-Analyse, HBA) beruht auf zwei Voraussetzungen, nämlich dass (a) das Studienprotokoll wissenschaftlich geeignet ist und (b) die Verwendung von (empfindungsfähigen) Tieren und die ihnen zugefügten Schäden für die Erreichung der Studienziele notwendig sind. Die 3R (Replace, Reduce, Refine) sind ein Leitprinzip für die Beurteilung, ob der Einsatz von Tieren, ihre Anzahl und die ihnen zugefügten Schäden notwendig sind. Kürzlich wurde ein ähnliches Leitprinzip zur Beurteilung, ob ein Studienprotokoll wissenschaftlich geeignet ist, vorgeschlagen – das 3V-Prinzip, das sich auf die drei Hauptaspekte der wissenschaftlichen Validität in der Tierforschung bezieht (Konstrukt-, interne und externe Validität). Hier analysieren wir die interne Kohärenz und Vereinbarkeit dieser beiden Prinzipien, gehen auf Konflikte innerhalb und zwischen den 3R und 3V ein und diskutieren ihre Implikationen für die Projektevaluierung. Wir zeigen, dass es einige wenige Konflikte und Kompromisse gibt, die aber durch geeignete Studiendesigns oder durch ethische Abwägung bei der HBA aufgelöst werden können. In Kombination bieten die 3V, 3R und die HBA somit einen kohärenten Rahmen für ein logisch strukturiertes Evaluationsverfahren zur Entscheidung über die Legitimität von Tierversuchsprojekten.

Resumen

Utilizar a animales para estudios de investigación provoca preocupaciones éticas que se tratan en las evaluaciones de proyectos en los que se ponderan los daños esperados a los animales frente a los beneficios previstos para la sociedad. Un análisis de daños y ventajas (HBA) depende de dos prerequisites: (a) el protocolo del estudio debe ser científicamente adecuado y (b) el uso de animales (sensibles) y los daños provocados a los mismos deben ser inevitables para conseguir los objetivos del estudio. Las 3 R (Reemplazar, Reducir y Refinar) ofrecen un principio rector para evaluar si el uso de animales, el número y los daños provocados son realmente necesarios. Recientemente se ha adoptado otro principio rector para evaluar si un protocolo de estudio es científicamente adecuado: el principio de las 3 V que hace referencia a los tres aspectos principales de validez científica en los estudios de investigación con animales (crear validez interna y externa). Aquí analizamos la consistencia interna y la compatibilidad de estos dos principios, además de tratar conflictos dentro y entre las 3 R y las 3 V, y debatimos sus implicaciones para la evaluación del proyecto. Mostramos que existen algunos conflictos y compensaciones, pero estos pueden resolverse mediante diseños de estudio adecuados o a través de una deliberación ética en el HBA. En combinación, las 3 R, las 3 V y el HBA ofrecen un marco coherente para llevar a cabo una evaluación estructurada de forma lógica para decidir la legitimidad de los proyectos de investigación con animales.